

Securing Birth Certificate Documents with DNA Profiles

Mark F. Tannian
Independent Researcher
tannianmf-hicss@yahoo.com

Christina Schweikert
St. John's University
schweikc@stjohns.edu

Ying Liu
St. John's University
liuy1@stjohns.edu

Abstract

The birth certificate is a document used by a person to obtain identification and licensing documents throughout their lifetime. For identity verification, the birth certificate provides limited information to support a person's claim of identity. Authentication to the birth certificate is strictly a matter of possession. DNA profiling is becoming a commodity analysis that can be done accurately in under two hours with little human intervention. The DNA profile is a superior biometric to add to a birth record because it is stable throughout a person's life and beyond. Acceptability of universal DNA profiling will depend heavily on privacy and safety concerns. This paper uses the U.S. FBI CODIS profile as a basis to discuss the effectiveness of DNA profiling and to provide a practical basis for a discussion of potential privacy and authenticity controls. As is discussed, adopting DNA profiles to improve document security should be done cautiously.

1. Introduction

A birth certificate is a document that enables a person to obtain genuine identification and licensing documentation years after birth. The birth certificate is referred to as a breeder document. The birth certificate provides limited information to an identity verifier. The birth certificate, as means for identity authentication, is less reliable than identification and licensing documents obtained later in life, such as a driver's license or passport. Unlike the passport or driver's license, historically the birth certificate has been a static document, meaning its contents are never renewed or refreshed.

Recently a birth certificate or certificate of live birth has been implemented as an electronic record [4] when submitted to an appropriate jurisdictional vital records agency by one of approximately 6,400 entities

that are authorized birth certificate issuers in the United States [10]. The electronic record contains public health related fields that are not commonly seen on the document typically associated with the term "birth certificate." A second document titled "Birth Certificate" or "Certificate of Live Birth" is what individuals typically consider the document of their birth. This second birth document is issued to individuals by a vital records agency, and primarily contains information including a child's name, birth date, birthplace, gender and parent information. Possession of the birth certificate is the dominant means of proving the identity on the certificate is that of the person presenting it or of a minor a guardian is assisting. The existence of two related documents produced by different issuers and for different purposes may be confusing. For purposes of clarity, the documentation submitted to the vital records agencies will be referred to as the certificate of live birth (CLB). The document issued by vital records agencies to individuals for identification purposes will be referred to as the birth certificate (BC). Historically, hospitals and midwives issued BCs to family members, but today in many cases, government entities are the legitimate issuers.

Biometrics within passports and identity documents has been incorporated [2]. The report titled "Birth Certificate Fraud" issued in 2000 by the Office of Inspector General of the U.S. Dept. of Health and Human Services suggests that biometrics be considered for incorporation in birth certificates [10]. The lack of progress on that front may be an issue of cost and technical capabilities. Another issue to consider is the change a person undergoes from birth to adulthood. There are few biometric identifiers that remain stable from birth to death. The DNA profile is the only biometric that remains stable from birth to death and beyond. The dead have been identified by tissue samples of an otherwise unrecognizable person or body part. Having the DNA profile established at birth allows for a person's life to have a definitive beginning for purposes of documentation and record keeping.

DNA analysis has become commonplace. It is being used for food safety and food counterfeiting

surveillance [23]. DNA analysis is being used for defecation attribution [18]. The web site ancestry.com is offering DNA analysis for \$89 and has tested more than 1.5 million people [6]. DNA analysis is quickly approaching commodity status. The FBI has approved a Rapid DNA Index System that will produce a Combined DNA Index System (CODIS) acceptable DNA profile within one to two hours from sample intake to result and requires no human intervention [7]. To date, the Rapid DNA system has not been approved for CODIS profile submission when operated from within a law enforcement booking station or agency. However, this federal program's technical achievement is an indication that DNA analysis can eventually be performed outside specialized laboratories.

This paper is an exploration of the potential convergence of DNA analysis, birth records management and identification document security. This convergence promises to increase document security significantly. The polymerase chain reaction (PCR) short tandem repeat (STR) DNA profiling has the potential to reliably link individuals to their birth records at time of birth and provide a reliable means of identity authentication many years later. This improved security, however, complicates birth certificate issuance and maintenance for years to come. Acceptability of universal DNA profiling will be contingent upon safeguards to protect personal privacy and safety. This paper raises some of the requirements that need to be considered and discusses the challenges they bring. Document security is a multi-level problem. The instance of the document must exhibit various integrity and authenticity properties. The system by which documents are issued and managed also requires careful consideration. This paper focuses on the individual birth record as opposed to document management systems.

This paper explores past efforts related to this topic and how DNA profiling could work in conjunction with document security, and closes with a discussion.

2. Past efforts

Numerous biometrics have been introduced to address a number of identification and security problems. Reliable individual recognition is critical to many processes that require accurate authorization and accountability. Biometric authentication has been automated to verify or recognize the identity of a living person based on a physiological or behavioral characteristic [16, 17, 24, 25]. A few better known physiological and behavioral characteristics currently used for automatic identification include DNA, fingerprints, voice, iris, retina, hand, face, handwriting,

keystroke, finger shape, as well as new measures, such as gait, ear shape, head resonance, optical skin reflectance and body odor. The ideal biometric characteristic has five qualities: robustness, distinctiveness, availability, accessibility and acceptability. The biometric should be: 1) unchanging on an individual over time ("robust"); 2) showing great variation over the population ("distinctive"); 3) that members of the population should have multiple instances of this measure ("available"); 4) ability to image or capture the measure's qualities using electronic sensors ("accessible"); 5) that people do not object to having this measurement taken of them ("acceptability") [16, 17, 19, 24]. Many uses of biometrics have focused on IT systems and facilities security [16, 17]. Performance based biometrics, such as keystroke or gait, are unrealistic as a means to authenticate an identity document given variety of physical environments in which identity documents are verified. The performance characteristics of an infant are a poor indicator of how that person in adulthood will perform (e.g. typing, walking). Fingerprints may be the most compelling option available today, but their size undergoes significant change from infancy to adulthood [3].

Methods to embed DNA information in identity documents have been proposed. Fuson applied for patents that document methods for including DNA on a birth certificate by either embedding an actual DNA sample or incorporating a chip containing DNA data [11, 12]. Other researchers have developed a printable ink that contains DNA for identification [15, 21]. These approaches, however, are potentially more costly and cumbersome than the technique presented here, which encodes and transforms CODIS loci information into a numerical value to be placed on the tangible certificate as well as in the online vital records entry. The transformation proposed in this article considers the privacy risk of capturing DNA information, which these past efforts do not adequately address. In terms of the efficiency of using DNA as a biometric, researchers at National Institute of Standards and Technology (NIST), are making progress on reducing the processing time for PCR-based STR markers [22].

3. Genetics primer

Within the nucleus of each human cell are 23 pairs of chromosomes. The mother contributes one chromosome to each pair and the other is from the father. Each chromosome is composed of deoxyribonucleic acid (DNA). DNA is a double helix structure that is designed to split into two strands during cell division. The two strands are joined

together by specialized chemical bonds between two organic compounds. This bonded pair of compounds is called a base pair. There are four distinct types of base compounds. (A - adenine, T - thymine, G - guanine, C - cytosine). Although the two strands are chemically different they are complementary in terms of information. This quality is derived by the special nature of the bond between adenine and thymine and the bond between guanine and cytosine. The catalog of base pairs is limited to A-T, T-A, G-C and C-G. This means a single strand of DNA dictates the composition of the double helix because the second strand must be composed of base compounds that will bond. This pairing behavior is exploited in DNA profiling.

The estimated number of base pairs among the 23 chromosome pairs is three billion [1]. Sequences of base pairs function collectively to act as a gene. A gene defines the chemical composition of proteins and other organic compounds necessary for life. Not all of the 3 billion base pairs appear to have an active part in cellular biology. Genes and non-functional sequences of base pairs are located on the same chromosome and same starting point along the double helix. Genes and non-functional sequences come in pairs, and are located at the same location of the same chromosome contributed by the mother and the chromosome provided by the father. The base compound composition within these genes and non-functional sequences can vary between maternal and paternal chromosomes. An allele is a specific base compound sequence of a gene or non-functional region. The term genotype commonly refers to a pair of alleles where each allele exists at the same specific location or locus on the maternal and paternal chromosome within a chromosome pair.

One composition dynamic within DNA is the presence of STRs. STRs are repetitions of short adjacent sequences consisting of base compounds, such as TATA or GTAGTA along a single DNA strand. The quantity of these tandem or adjacent compound sequence repetitions defines the length of the allele. For example, the allele designations for D3S1358 shown in Table 1 signify length as the number of times a particular expected base combination sequence (i.e. AGAT or TCTA depending on strand [9]) repeats starting at the locus D3S1358. An allele designated as 15 is one where 15 repetitions of the expected base sequence occurred. An allele with a decimal point value indicates the degree of completion of the expected sequence within the last repetition. A locus designator (e.g. D3S1358) starting with D indicates the chromosome (e.g. chromosome 3) and identifies a unique DNA segment (e.g. 1358) along the chromosome. The other locus label conventions

seen in Table 1 follow historical naming that requires a reference lookup to determine the actual location.

Although each chromosome pair consists of a contribution from the mother and father, the chromosome's contributor cannot be attributed by general characteristics such as length, orientation, color, location, weight or shape. A reference chromosome from the mother or father is needed for partial sequencing and compared to the partial sequencing of the offspring's chromosome in order to attribute origin. This is done in paternity cases, which is not the purpose of the CODIS. In the criminal DNA forensics context, attributing the parental origin of each of the 46 chromosomes appears to be unnecessary [20]. Genotype notation is not an ordered list. The genotype for D3S1358 of allele of 15, 16 is equivalent to 16, 15. The parental origin of the particular DNA strand containing the allele of 15 at D3S1358 is not determined. During the PCR analysis process, DNA strands are snipped apart bio-chemically and the process does not maintain the information of which pair member contributed to which snippet. This property is utilized in the discussion regarding privacy.

4. Document security

When attempting to improve the reliability of linkage between a person and their identity documentation, the challenge is not identification (seeking to know who an individual is from a population), but a challenge of authentication or verification (is the claimant truly the person identified by document). Authentication requires that a person be enrolled in order to initialize the identity and submit a means by which this person will support their claim of identity in the future.

Biometrics, in the context of identity documents, have usage dynamics that are different from systems security. The CLB is issued by a loosely coordinated group of issuers, and the verifiers of identity are unlikely to be those who issued the BC. For example, a baby born in New York City (NYC) will have a birth certificate issued by the NYC Department of Health, and a verifier could be the Nebraska Department of Motor Vehicles. The international dimension of issuance and verification raises complexity even further. Decentralized issuance and uncoordinated verification lead to dynamics unlike those experienced within a closed system.

The identity document issuer and verifier are concerned about the document's accuracy, authenticity and integrity. The verifier needs a means to authenticate the document's legitimacy as well as the claimant-to-document relationship. Birth records without biometrics have fewer privacy implications for

the individual being identified. Affixing valuable immutable data such as DNA sequencing raises the privacy risk a person experiences. The next sections explore how DNA can be utilized for authentication, propose a means to achieve privacy and suggest a means to maintain authenticity and integrity.

4.1. Individual authentication

When considering the use of DNA to provide authentication, the question being asked is “How can DNA analysis be used to ensure the person claiming a birth certificate is theirs is being honest?” Underlying this question are questions related to the process of DNA analysis, process reliability, results interpretation and the chances and types of error.

DNA use in forensics has primarily been used to aid the justice system in criminal and civil cases. The notion of relating evidence (ex. blood sample, child with disputed paternity) with individuals is conceptually similar to relating a documented DNA profile bound to an identity document to the person who claims to be represented by that document. DNA forensics analysis involves one or both of two dominant methods. The two processes are the variable numbers of tandem repeats (VNTR) process and the PCR process. The PCR process depends on STRs and is agile in terms of method acceleration and is highly precise in its results. Disadvantages related to PCR-based STR methods are that the individuating power per locus is lower and contamination during the process has an exaggerated effect. The FBI uses the PCR-based STR process to populate CODIS, which is a collection of DNA profiles used in law enforcement. As of January 2017, the system will use 20 core STR loci, which is seven more loci than in DNA profiles that exist prior to this date [14].

In order for CODIS to be successful as a law enforcement and prosecutorial tool, the scientific basis of the PCR-based STR profile utilized has undergone significant review in the areas of genetics, reliability, repeatability, population genetics and statistics. Adoption of CODIS profiles as the basis for DNA authentication of birth certificates is a reasonable choice. Although the newest CODIS profile of 20 STR loci has not been instituted operationally, it is formally adopted. As long as the sample contains cells from one DNA contributor, the PCR-based STR process produces an unambiguous profile. However, there remains the question of whether a fraudulent claimant will be able to successfully authenticate (false positive) or whether the proper claimant could be unsuccessful in proving his/her claim (false negative).

In order to make these concepts more tangible, a scenario of a fictional person called Jo is explored. Jo

is gender neutral. The core CODIS loci are autosomes, meaning the loci are not located on the X-Y gender chromosomes. Table 1 shows Jo’s CODIS DNA profile.

Table 1: Jo’s CODIS DNA profile - compiled by drawing from allele types and frequencies published in the “Caucasian 2015 Expanded FBI STR Loci Allele Frequencies” [8].

Locus	Genotype	Allele Frequencies
D3S1358	15, 16	0.2475, 0.2327
vWA	17, 18	0.2673, 0.2178
D16S539	12, 11	0.3416, 0.2723
CSF1PO	12, 11	0.3267, 0.2995
TPOX	8, 11	0.5470, 0.2550
D8S1179	13, 14	0.3342, 0.2054
D21S11	30, 29	0.2327, 0.1807
D18S51	14, 17	0.1757, 0.1535
D2S441	11, 14	0.3094, 0.2624
D19S433	14, 13	0.3490, 0.2797
TH01	9.3, 6	0.3045, 0.2252
FGA	22, 21	0.1881, 0.1757
D22S1045	15, 16	0.3639, 0.3168
D5S818	11, 12	0.4084, 0.3515
D13S317	11, 12	0.3119, 0.3094
D7S820	10, 11	0.2896, 0.2030
D10S1248	13, 14	0.3366, 0.2748
D1S1656	17.3, 15	0.1510, 0.1436
D12S391	18, 21	0.1757, 0.1337
D2S1338	17, 19	0.1931, 0.1510

Jo is as “common” as a Caucasian person can be with the CODIS profile. Jo’s profile shows the discriminating power of the CODIS scheme by assembling a profile consisting of the two most common alleles for each locus. At first glance it is apparent that even the most common alleles for some loci are not so common (ex. D18S51, D12S391). Jo is the offspring of parents who happen to contribute the most common alleles of each locus for Caucasians within the sample used to produce the FBI reference statistics table for Caucasians. A somewhat unusual characteristic of this profile is that all genotypes are heterozygous (the alleles contributed by the parents are different). It would be reasonable to see at least one genotype that is homozygous (the same allele contributed by both parents). These genotype distinctions influence the statistical calculations. As will be shown, homozygous genotype occurrence at a locus is less frequent within a population and therefore makes a profile more distinctive and less common.

According to the National Research Council (NRC) report [1] recommendations adopted by the FBI in the FBI Quality Assurance Standards document [5],

the following are the common computations related to DNA profiles.

Random-Match Probabilities:

Equation 1.A: $p^2 + p(1 - p)\theta$ (homozygous)

Equation 1.B: $2p_i p_j$ (heterozygotes)

The letter p represents probability of an allele occurring within a population, which the FBI are calling frequency in their tables. The expression that follows the squaring of p in equation 1.A is a correction factor to adjust for the underlying assumptions of a random mating population. Reality is that females and males do not routinely seek mates from random places (e.g. mother from Maine and father from Oklahoma) within the U.S. Convenience is a likely factor in most pairings and location and demographic factors influence just how random genotypes will be. There is a strong correlation of homozygosity and the degree of familial relation between mates, and isolated communities are likely to have individuals with a higher degree of homozygosity than what would occur if truly random mating were to occur. Population groupings like Caucasian, African American, and American Indian influence the frequencies of occurrence of alleles within genotypes and proper statistical calculations require that the population group of the individual from which a profile originates be considered. The θ is a constant that addresses the degree of isolation of the community from which an individual originates. The θ value is recommended by the NRC to range from 0.01 to 0.03 of which 0.03 is considered highly conservative from a false positive point of view. This is due to θ increasing the probability of occurrence of a genotype, which in turn suggests the genotype is less discriminating in a particular circumstance.

Equation 1.B is fairly straightforward. The reason for the doubling of the result is that the allele frequency is independent of whether the allele source was the mother or father. There are two opportunities for the genotype to occur. The subscripts i and j refer to a specific allele of a locus. Because 1.B is for heterozygote genotypes, the alleles are not the same and therefore i and j cannot be equal. Applying 1.B to Jo's profile and for the particular locus vWA, the equation would be written as $2p_{17}p_{18} = 2 * 0.2676 * 0.2178 = 0.1152$.

Equations 1.A and 1.B compute the probability that a particular genotype occurs at a specific locus. The statistical theory supporting this and other computations depend on the loci being gender neutral and not subject to natural selection.

A DNA profile consists of multiple loci, and the discriminating power of the profile results from being

able to combine the probabilities or frequencies of the observed genotypes to determine the uniqueness of the overall profile. The product rule is the recommended calculation of profile frequency [1]. This calculation depends on the notion of linkage equilibrium among the loci, which is the population genetics concept that over time generations within populations will mate in such a way the genotypes that occur in selection neutral locations will occur independent of each other, which allows us to treat the loci as independent random variables necessary for the product rule to be valid. Therefore a DNA profile's frequency can be determined by applying 1.A or 1.B to each observed genotype depending on its homozygosity and multiplying each genotype probability result together using the product rule.

In the case of Jo, the profile frequency is 8.0493×10^{-19} . This value is the match probability, which measures the probability of this profile being presented by someone else (i.e. false positive). These other possible people are assumed to be random people within the population, but not relatives. The most challenging source of false positives is Jo's relatives. An identical twin is able to successfully authenticate. Existence of identical twins should be documented on the birth certificate, and other authentication controls will be needed for that situation. Jo's parents, uncles/aunts, nieces/nephews have match probabilities relative to Jo's profile that are significantly higher. The match probability of a parent, uncle/aunt, first cousin matching Jo's COIDS profile, consisting of the two most common alleles for each locus, are 0.1351 (a conservative estimate), 0.0336, 0.0161 respectively. A distant unilineal relative of Jo's has a match probability of at least 0.0074. The formulae used for these match probability calculations can be found in the NRC report [1].

The birthing facility would be the appropriate party to perform DNA profile enrollment. They have physical possession of the infant and often have witnessed the child's birth. They also have the medically trained staff to obtain and process a DNA sample, which in time should require little laboratory training. Infants born at home or in non-institutional settings would need to be brought into an authorized facility for a medical exam and formal birth registration. The resulting DNA profile would be translated into a data structure that addresses integrity, authenticity and privacy requirements prior to CLB submission to the vital records agency. The CLB issuer should be the entity that provides assurances of the DNA profile's authenticity, because they performed the DNA profile enrollment. The BC issuer assures authenticity over the entire BC, but the CLB issuer performed the enrolling DNA profile protocol.

4.2. Privacy

In this discussion of privacy protection the most significant assumption being made is that the DNA profile be composed of PCR-based STRs or be performed with a profiling protocol that characterizes a person's alleles with equal or better precision and reliability.

With recorded lifespans occasionally exceeding 120 years, a solution involving DNA privacy should target at least 150 years of protection. Privacy protections would need to protect a person's privacy from the first minutes after birth when the birth is documented to their death. Individuals are born nearly every second of every day. The privacy time clock is reset for each of these births. A birth record management system will need to institute new protections regularly in order to provide 150 years of privacy protection for each subsequent newborn. A privacy concern is that genetic relatives are a threat to each other in terms of DNA privacy. An older birth record for a grandparent may be a privacy concern for their grandchildren. Having the DNA profile directly readable on the birth certificate raises significant privacy concerns.

Although the PCR STR loci are chosen to be ideally non-functioning regions within chromosomes, some loci are related to genes. Often loci are selected from regions of chromosomes that are not relevant genetically to diseases or subject to natural selection [1, 13].

Using the FBI profiling schema should avoid medically sensitive portions of the DNA. This is a positive privacy accommodation that reduces medical privacy threats to the profile data. However, the immutable nature of DNA and the similarity among relatives raises the concern that the profile in unscrupulous hands would be used maliciously against the individual or a family member or an entity trusting the validity of the DNA authentication process.

Although statistics are required to determine the significance of a match, statistics and judgment are not used to determine the profile and evaluate whether a match has occurred. The precision of the PCR-based STR process allows for exact determination of the genotype at each locus each time it is properly utilized. Unlike other biometrics, DNA profile determination is consistent and controlled. The final results are highly repeatable, and therefore the match determination is a simple logical evaluation. The authentication evaluation is, "Do both the documented profile and provided profile contain the same genotype at each loci?" A PCR-based STR profile match is accepted or rejected based on complete agreement between two

profiles. The DNA profile is an immutable and innate passphrase every person possesses.

A requirement for privacy protections is that they can complicate but must not prevent effective authentication and authenticity verification. In order to provide privacy for this "bio-phrase," one can take a page from the history of passphrase protection. One can construct a DNA profile and encode it into a structured profile data unit (PDU), and perform one-way cryptographic transformations on that PDU that produces a unique result. These transformations limit the ability to restore and determine the contents of the original PDU, which contains the DNA profile, while preserving the distinctiveness of the DNA profile. Cryptographic hash functions like SHA-2 or SHA-3 are algorithms that suffice today. By avoiding a reversible process, such as encryption, there is no need to rely on secrecy of decryption keys to maintain privacy protection. However, another page in passphrase history is the offline attack on passphrases. The equivalent attack would be for an attacker to gain access to a birth record and attempt to determine the DNA profile by postulating numerous, possibly millions, DNA profiles and transforming them until an attacker's attempts result in a match.

Returning to Jo's profile, a simple concatenation of the genotypes would allow us to construct a bio-phrase without losing profile information. Prior to applying cryptographic algorithms the genotype data would undergo simple pre-processing. First step is to sort the locus allele pairs to force consistency of placement of allele values within the PDU. Next is to normalize them into three digit integers by multiplying each allele value by 10. Each result under 100 would have a zero prepended. This process avoids parsing ambiguity within the PDU contents. Finally, concatenating the normalized allele values to form genotype sub-strings. Table 2 shows results of the encoding steps in context of Jo's DNA profile. There is an opportunity to improve privacy by randomly concatenating the genotype sub-strings when forming the PDU. The order of loci provided in Table 1 is not significant to the interpretation of the DNA profile. By randomly ordering the profile loci, 20! possible genotype sub-string orderings are introduced. Pre-computation attacks, such as rainbow tables, will be more difficult to execute. The order of the genotypes would be documented and accessible to the verifier in order to assemble the genotype sub-strings into the correct order during verification.

It is reasonable to anticipate that different DNA profiling protocols will be introduced over time and ambiguity regarding the protocol must be avoided when verifying an individual's BC. The DNA profile protocol identifier, loci order pattern and CLB record

identifier would be among the accessible fields of the BC in order for a verifier to reproduce the PDU generated during enrollment. The genotype ordering pattern metadata, CLB record identifier, DNA profiling protocol identifier would also be incorporated as part of the PDU thus ensuring the loci pattern, correct CLB record and profiling process are used or referenced during verification. The PDU is hashed using a cryptographic hash function. The result is submitted along with other CLB fields to the appropriate vital records agency.

Table 2: Encoding of Jo's profile

	Profile		Sorted		Genotype Sub-string
	Result		Long	Short	
D3S1358	15	16	16	15	160150
vWA	17	18	18	17	180170
D16S539	12	11	12	11	120110
CSF1PO	12	11	12	11	120110
TPOX	8	11	11	8	110080
D8S1179	13	14	14	13	140130
D21S11	30	29	30	29	300290
D18S51	14	17	17	14	170140
D2S441	11	14	14	11	140110
D19S433	14	13	14	13	140130
TH01	9.3	6	9.3	6	093060
FGA	22	21	22	21	220210
D22S1045	15	16	16	15	160150
D5S818	11	12	12	11	120110
D13S317	11	12	12	11	120110
D7S820	10	11	11	10	110100
D10S1248	13	14	14	13	140130
D1S1656	17.3	15	17.3	15	173150
D12S391	18	21	21	18	210180
D2S1338	17	19	19	17	190170

4.3. Integrity and authenticity

Integrity and authenticity of document content can be assured through specialized documentation materials and printing processes used to issue a BC. Integrity and authenticity of the electronic CLB will require digital security controls. Similar digital data security controls will prove useful for digital data stored on the BC. The integrity of all of the various

CLB fields including the DNA profile (contained within the protected PDU) and supporting profile fields must be verifiable as an ensemble within the CLB and BC. Performing a cryptographic checksum incorporating all of the birth record (CLB or BC) data fields can enable record integrity verification. However, the cryptographic checksum could be substituted with a new value that incorporates fraudulent changes. To prevent substitution of the checksum, data authenticity services are required. The CLB issuer will authenticate the protected PDU as well as authenticate the overall CLB. The distinct authentication on the protected PDU is necessary for BC content validation.

Anticipating coordination among CLB submitting institutions, vital records agencies and BC validators is not practical. Therefore, authenticating BC content cannot be dependent upon shared secrets. Digital signatures can provide the authenticity services needed. This will complicate BCs further, because the public-key certificates of both the CLB submitter and the vital records agency must be accessible to the verifier in order to validate the digital signatures covering BC content.

The introduction of public-key certificates leads directly to the requirement of a public key infrastructure. Public-key certificates need to be issued by an entity each validator can trust. There are more than 6,400 authorized CLB submitters in the U.S. and those organizations that grant CLB submission authorization would be the natural parties to issue the CLB submitter's public-key certificate. For these purposes, the BC validator would be better served by a flat certificate authority hierarchy thus improving the BC's ability to be self-contained. A validating apparatus could be configured with a managed repository of trusted public-key certificates for each certificate authority that may have signed the CLB submitter and BC issuer public-key certificates.

Adding DNA profile information and related information will increase the amount of accessible information significantly. Digital signatures would complement printed seals, and will be easier for a verification apparatus to accurately authenticate. A contactless storage device integrated into the certificate material should store all BC fields and any additional information not visible on the BC. Printing protected PDU information on the BC should be considered carefully. Visual BC replicas will pose a threat to the protected PDU in the long term. Copies of the BC will not update in the event that privacy and authenticity protections on the official BC document and digital content are refreshed.

4.4. Summary of services

The binding between the individual and their BC is provided by a DNA profile produced using a validated DNA profiling protocol. The genotypes are randomly organized and concatenated with additional context information to form a PDU that is cryptographically hashed. The CLB issuer digitally signs the protected PDU. The BC issuer will incorporate the protected PDU and corresponding CLB digital signature within an issued BC. Public-key certificates for the CLB and BC issuers will be incorporated with BC content in order for the verifier to validate the BC's contents authenticity. The BC issuer will digitally sign across all of the BC's contents.

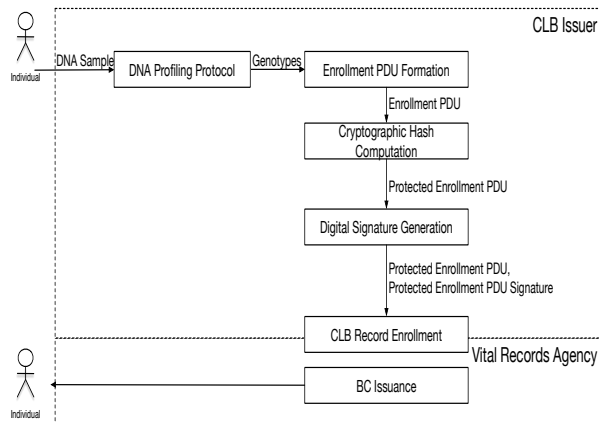


Figure 1: Depiction of DNA profile generation and protection

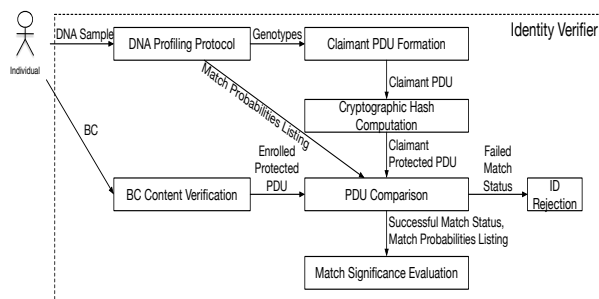


Figure 2: Depiction of DNA profile verification

At identity verification, BC contents are authenticated and the claimant's DNA profile is produced using the same profiling protocol used at enrollment. A claimant version of the protected PDU is generated and compared to the enrolled version. If a

match does not occur, the claimant is not the same person enrolled on the document or a profiling protocol error occurred. If a match occurs, without revealing the submitted DNA profile, a verifier's profiling apparatus will produce match probabilities using the claimant's DNA profile and current population statistics for use in match significance determination.

5. Discussion

The CODIS profile was selected as a potential DNA profiling protocol because it has an established scientific pedigree on which to base this discussion. A more sensitive identity-profiling protocol could be introduced and validated with the same rigor as CODIS profiles. Identifying and the preserving the parental origin of the alleles may improve DNA profile reliability, but the reference-parent's privacy, likely the birth mother, is at additional risk. If a substitute protocol produces a repeatable sequence of discriminating information that undergoes a literal value match, the privacy suggestions discussed in this paper may apply.

DNA based document authentication has great potential to detect fraudulent claims by genetic strangers attempting to impersonate. Relatives pose a comparatively high false positive risk. Prosecutors introduce additional evidence to rule out relatives being involved in a crime. Requiring additional documents may not be an option for a verifier because the BC may be the only legitimate document available for identity verification. Verifiers can avoid some fraud attempts by determining that age and gender of the claimant is consistent with what is documented on the BC. Verifiers would need to accept that there is a risk of a relative successfully posing as the claimed identity. Assuming CODIS profiles are adopted, research may be needed to determine how common "CODIS twins" (relatives with matching CODIS profiles) are within the U.S. population. This would aid in determining the fraud potential posed by relatives.

International adoption of DNA enhanced birth records is anticipated to be staggered over time due to factors including costs, priorities, societal values and logistical complexity. Traditional visible contents of the BC on enhanced BCs can be used by jurisdictions not able to validate available DNA enhanced data. Those claimants without DNA enhanced birth certificates will not be able to provide a higher level of identity assurance, but "DNA enhanced" jurisdictions will likely rely on the policies they have today for accepting and performing identification using foreign BCs. Many nations will want to influence the choice of DNA profiling protocol they utilize. A United

Nations organization similar to the International Civil Aviation Organization (ICAO) may be the proper forum to negotiate details in the areas like DNA profiling, PDU structure, contactless storage and cryptography. Each chosen DNA profiling protocol will need corresponding statistical analyses of the ethnic populations residing in the nations that adopt the protocol, which will need to be shared across jurisdictions. International standards and agreements will be needed to avoid an overabundance of DNA enhancement approaches that would make international compatibility costly and technically challenging.

Failure to verify a legitimate claimant (false negative) results from contamination or test execution errors. An authoritative report recommends independent re-testing to address false negative concerns [1]. The PCR process involves amplifying the quantity of alleles at each locus making it highly unlikely the alleles are mischaracterized. This precision nearly guarantees that if a true match is possible that match will occur. False negatives using PCR-based STR should be expected to be rare using properly executed procedures. Genetic chimerism is one of possibly other rare naturally occurring genetic conditions that could result in a false negative.

DNA profiling has the potential to upgrade the legitimacy of birth certificates, and improve the reliability of documents such as passports, visas, school records and license documents. Missing person records could be populated with DNA enhanced birth record information drawn from the CLB or BC. Loved ones could find closure from accidents that may have significantly damaged human remains. DNA analysis is not novel in these use cases, but to have a reliable, comprehensive collection of DNA enhanced birth records may improve identification in these situations. An open-ended victim identity search with limited search constraints will prove to be difficult because of the privacy protections on the CLB records, and the variation of DNA profiling protocols used historically may require several profiles to be generated. With sufficient access, resources and time, the victim's identity could be determined.

DNA profiles are fundamentally immutable and are reusable authenticators. This combination puts pressure on DNA privacy. Identity thieves need not resort to cryptanalysis, but can obtain DNA samples that are left behind in daily routines or at times of enrollment or verification. Record level security cannot address these threats, but these threats should be considered systematically when evaluating DNA profiles as a document security feature.

A 150-year security requirement for protecting a historically static low privacy risk record will unsettle birth record management. Adding protected DNA

profiles will introduce digital authenticity and privacy requirements that must be met under threats that will span over a century or more for every registered person. Cryptanalysis techniques and computational power will improve over time. Public-key certificates are commonly set to expire in time periods under a century to minimize risk from lost private keys or weakened cryptographic algorithms. One-way transformations may leak more original information with improved cryptanalysis. Contactless storage device lifetimes will need to be extended or be periodically replaced securely. DNA profiling protocols will change over time causing support for legacy DNA profiles to be burdensome for verifiers. These factors challenge the low maintenance tradition of birth record management. It may be necessary for CLBs and BCs to undergo a periodic refresh in order to utilize an updated DNA profile protocol and improve data protections. A cost benefit analysis performed today may show DNA profiling unattractive, but as DNA analysis becomes routine, policy-makers should not lose sight of the privacy consequences. Securing birth records with DNA profiles appears to be within reach, but should be done cautiously.

6. Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments. We would like to thank Joseph Idziorek for his early encouragement in pursuing this line of research. We would also like to thank Maura Flannery for her molecular biology guidance.

7. References

- [1] *The Evaluation of Forensic DNA Evidence*, National Academy Press, 1996, p. 272.
- [2] *Ensuring the Security of America's Borders Through the Use of Biometric Passports and Other Identity Documents*, 109-24, U.S. House of Representatives, 2005.
- [3] "Biometrics / Advantages and disadvantages of technologies," 2006; http://biometrics.pbworks.com/w/page/14811349/Advantages_and_disadvantages_of_technologies.
- [4] "The Birth Certificate (Finally) Goes National," 2014; http://www.cdc.gov/nchs/features/birth_certificate_goes_final.htm.
- [5] "Quality Assurance Standards For Forensic DNA Testing Laboratories," 2016; <http://www.fbi.gov/about->

[us/lab/biometric-analysis/codis/qas-standards-for-forensic-dna-testing-laboratories-effective-9-1-2011](http://www.fbi.gov/about-us/lab/biometric-analysis/codis/qas-standards-for-forensic-dna-testing-laboratories-effective-9-1-2011).

[6] "DNA Test for Ethnicity and Genelogical DNA Testing," 2016; <http://dna.ancestry.com/>.

[7] "Rapid DNA Analysis," 2016; <http://www.fbi.gov/about-us/lab/biometric-analysis/codis/rapid-dna-analysis>.

[8] "2015 FBI Population Data for the expanded CODIS core STR loci," 2016; <http://www.fbi.gov/about-us/lab/biometric-analysis/codis/expanded-fbi-str-2015-final-6-16-15.pdf>.

[9] "STR Fact Sheet - D3S1358," 2016; http://www.cstl.nist.gov/strbase/str_D3S1358.htm.

[10] J.G. Brown, *Birth Certificate Fraud*, U.S. Dept. of Health and Human Services, 2000.

[11] L. Fuson, *DNA Enabled Certificate*, 20110316267, USPTO, 2011.

[12] L. Fuson, *Biometric Birth Certificate*, 20130020793, USPTO, 2013.

[13] D.R. Hares, "Expanding the CODIS core loci in the United States," *Forensic Science International: Genetics*, vol. 6, no. 1, 2012, pp. e52-e54; DOI <http://dx.doi.org/10.1016/j.fsigen.2011.04.012>.

[14] D.R. Hares, "Selection and implementation of expanded CODIS core loci in the United States," *Forensic Science International: Genetics*, vol. 17, 2015, pp. 33-34; DOI <http://dx.doi.org/10.1016/j.fsigen.2015.03.006>.

[15] M. Hashiyada, "Development of Biometric DNA Ink for Authentication Security," *The Tohoku Journal of Experimental Medicine*, vol. 204, no. 2, 2004, pp. 109-117; DOI 10.1620/tjem.204.109.

[16] A. Jain, L. Hong and S. Pankanti, "Biometric identification," *Commun. ACM*, vol. 43, no. 2, 2000, pp. 90-98; DOI 10.1145/328236.328110.

[17] A.K. Jain, A. Ross and S. Prabhakar, "An introduction to biometric recognition," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 14, no. 1, 2004, pp. 4-20; DOI 10.1109/TCSVT.2003.818349.

[18] G. Kolata. "Devious Defecator' Case Tests Genetics Law." New York Times, June 2, 2015, D6.

[19] V. Matyáš and Z. Říha, "Biometric authentication — security and usability," *Advanced Communications and Multimedia Security*, Springer, 2002, pp. 227-239.

[20] K. Norrgard, "Forensics, DNA fingerprinting, and CODIS," *Nature Education*, vol. 1, no. 1, 2008, pp. 35.

[21] A. Oshima and F. Yusa, *DNA-Containing Ink Composition*, 20110229881, to Nagahama Bio-Laboratory Inc., USPTO, 2011.

[22] E.L. Romsos and P.M. Vallone, "Rapid PCR of STR markers: Applications to human identification," *Forensic Science International: Genetics*, vol. 18, 2015, pp. 90-99; DOI <http://dx.doi.org/10.1016/j.fsigen.2015.04.008>.

[23] K. Stabiner. "Private Eyes in the Grocery Aisles." New York Times, April 5, 2015, BU1.

[24] J.L. Wayman, A. Jain, D. Maltoni, et al., *Biometric Systems: Technology, Design and Performance Evaluation*, Springer London, 2010.

[25] A.C. Weaver, "Biometric authentication," *Computer*, vol. 39, no. 2, 2006, pp. 96-97; DOI 10.1109/MC.2006.47.